

Why is it hard to beat $O(n^2)$ for Longest Common Weakly Increasing Subsequence?

Adam Polak*

Department of Theoretical Computer Science
Faculty of Mathematics and Computer Science

Jagiellonian University
`polak@tcs.uj.edu.pl`

Abstract

The Longest Common Weakly Increasing Subsequence problem (LCWIS) is a variant of the classic Longest Common Subsequence problem (LCS). Both problems can be solved with simple quadratic time algorithms. A recent line of research led to a number of matching conditional lower bounds for LCS and other related problems. However, the status of LCWIS remained open.

In this paper we show that LCWIS cannot be solved in $O(n^{2-\varepsilon})$ time unless the Strong Exponential Time Hypothesis (SETH) is false.

The ideas which we developed can also be used to obtain a lower bound based on a safer assumption of NC-SETH, i.e. a version of SETH which talks about NC circuits instead of less expressive CNF formulas.

*This work was supported by the Polish Ministry of Science and Higher Education program *Diaamentowy Grant*.

1 Introduction

Despite attracting interest of many researches, both from theoretical computer science and computational biology communities, for many years the classic Longest Common Subsequence problem (LCS) have not seen any significant improvement over the simple $O(n^2)$ dynamic programming algorithm. The current fastest, $O(n^2/\log^2 n)$ algorithm by Masek and Paterson [8], dates back to 1980.

Difficulties in making progress on the LCS inspired studying numerous related problems, among them the Longest Common Increasing Subsequence problem (LCIS), for which Yang, Huang, and Chao [10] found a quadratic time dynamic programming algorithm. Their algorithm was later improved by Sakai [9] to work in linear space. Even though both these algorithms are devised to compute the Longest Common Increasing Subsequence, they can be easily modified to compute the Longest Common Weakly Increasing Subsequence (LCWIS). The latter problem, first introduced by Kutz et al. [7], can be solved in linear time in the special case of a 3-symbols alphabet, as proposed by Duraj [6]. However, despite some attempts over the last decade, no subquadratic time algorithm has been found for the general case of LCWIS.

A recent line of research led to a number of conditional lower bounds for polynomial time solvable problems. In particular Abboud, Backurs, and Vassilevska Williams [1], and independently Bringmann and Künnemann [5] proved that LCS cannot be solved in $O(n^{2-\epsilon})$ time unless the Strong Exponential Time Hypothesis (SETH) is false.

Hypothesis 1 (Strong Exponential Time Hypothesis). *There is no $\epsilon > 0$ such that for all $k \geq 3$, k -SAT on N variables can be solved in $O(2^{(1-\epsilon)N})$ time.*

Moreover, Bringmann and Künnemann [5] proposed a general framework for proving quadratic time hardness of sequence similarity measures. Within this framework, it is sufficient to show that a similarity measure *admits an alignment gadget* to prove that this similarity measure cannot be computed in $O(n^{2-\epsilon})$ time unless SETH is false. Besides LCS many other similarity measures, e.g. Edit Distance and Dynamic Time Warping, fall into this framework. However, it seems that neither LCIS nor LCWIS admits an alignment gadget.

In this paper we show that LCWIS cannot be solved in $O(n^{2-\epsilon})$ time unless SETH is false. We do this by proving the following theorem.

Theorem 2. *If the Longest Common Weakly Increasing Subsequence problem for two sequences of length n can be solved in $O(n^{2-\epsilon})$ time, then given a CNF formula on N variables and M clauses it is possible to compute the maximum number of satisfiable clauses (MAX-CNF-SAT) in $O(2^{(1-\epsilon/2)N} \text{poly}(M))$ time.*

Our reduction is modelled after previous hardness results based on SETH, in particular [1] and [4]. We go through the Most-Orthogonal Vectors problem, and construct *vector gadgets* such that two vector gadgets have large LCWIS iff the corresponding vectors have small inner product. The crucial ingredient is a construction that lets us combine many vector gadgets into two sequences such that their LCWIS depends on the largest LCWIS among all pairs of vector gadgets.

Unlike $P \neq NP$ and several other common assumptions for conditional lower bounds in computational complexity, SETH is considered by many not a very safe working hypothesis. Recently, Abboud et al. [2] came up with a weaker assumption, which still allows to prove many previous SETH-based lower bounds. More specifically, they propose a reduction from satisfiability of *non-deterministic Branching Programs* [3] (BP-SAT) to LCS and any other similarity measure which admits an alignment gadget. Their reduction implies that the existence of a strongly subquadratic

time algorithm for LCS not only refutes SETH but also has stronger consequences, e.g. the existence of an $O(2^{(1-\delta)N})$ time algorithm for satisfiability of NC circuits. For an in-depth discussion of consequences of their reduction and motivations to study such reductions please refer to the original paper [2]. At the end of our paper we briefly explain how the ideas which we developed can be used to obtain a reduction from BP-SAT to LCWIS.

Unfortunately, our techniques are not sufficient to prove a similar lower bounds for LCIS.

2 Preliminaries

Let us start with the formal definition of the LCWIS problem.

Definition 3 (Longest Common Weakly Increasing Subsequence). *Given two sequences A and B over an alphabet Σ with a linear order \leq_Σ , the Longest Common Weakly Increasing Subsequence problem asks to find a sequence C such that*

- *it is weakly increasing with respect to \leq_Σ ,*
- *it is a subsequence of both A and B ,*
- *and its length is maximum possible.*

We denote the length of C by $\text{LCWIS}(A, B)$.

For example, $\text{LCWIS}(\langle 1, 2, 5, 2, 5, 3 \rangle, \langle 2, 4, 5, 2, 3, 4 \rangle) = 3$, and the optimal subsequence is $\langle 2, 2, 3 \rangle$.

To simplify further arguments we introduce, as an auxiliary problem, the weighted version of LCWIS.

Definition 4 (Weighted Longest Common Weakly Increasing Subsequence). *Given two sequences A and B over an alphabet Σ with a linear order \leq_Σ and the weight function $w : \Sigma \rightarrow \mathbb{N}_+$, the Weighted Longest Common Weakly Increasing Subsequence (WLCWIS) problem asks to find a sequence C such that*

- *it is weakly increasing with respect to \leq_Σ ,*
- *it is a subsequence of both A and B ,*
- *and its total weight, i.e. $\sum_{i=1}^{|C|} w(C_i)$, is maximum possible.*

We denote the total weight of C by $\text{WLCWIS}(A, B)$.

Lemma 5. *For a sequence $X = \langle X_1, X_2, \dots, X_{|X|} \rangle$ let \hat{X} denote a sequence obtained from X by replacing each symbol a by its $w(a)$ many copies, i.e.*

$$\hat{X} = X_1^{w(X_1)} X_2^{w(X_2)} \dots X_{|X|}^{w(X_{|X|})}.$$

Computing the WLCWIS of two sequences A and B of total weight at most n can be reduced to computing the LCWIS of two sequences \hat{A} and \hat{B} of length at most n .

Proof. We have to show that $\text{WLCWIS}(A, B) = \text{LCWIS}(\hat{A}, \hat{B})$. Every common weakly increasing subsequence C of A and B translates to a common weakly increasing subsequence \hat{C} of \hat{A} and \hat{B} , and the length of \hat{C} equals the total weight of C , thus $\text{WLCWIS}(A, B) \leq \text{LCWIS}(\hat{A}, \hat{B})$.

The proof of inequality $\text{WLCWIS}(A, B) \geq \text{LCWIS}(\hat{A}, \hat{B})$ is by induction on the alphabet size. Denote by C the longest common weakly increasing subsequence of \hat{A} and \hat{B} , by a the smallest

symbol in the alphabet, and by k the number of occurrences of symbol a at the beginning of C . Note that k must be a multiple of $w(a)$, because otherwise C could be extended. To finish the proof, cut off the shortest prefixes of A and B containing at least $k/w(a)$ occurrences of symbol a , remove all symbols a from remaining suffixes, and apply inductive hypothesis to get a sequence that can be appended to $a^{k/w(a)}$ to get a common weakly increasing subsequence of A and B of the desired total weight. \square

The curious reader looking for a more detailed argument is referred to [1], where an analogous lemma for the weighted version of LCS is presented. Its proof can be directly translated to another proof for the case of WLCWIS.

3 Reduction from MAX-CNF-SAT to LCWIS

This section is devoted to proving Theorem 2. We do this by showing a reduction from the Most-Orthogonal Vectors problem, introduced by Abboud, Backurs, and Vassilevska Williams [1].

Definition 6 (Most-Orthogonal Vectors). *Given two sets of vectors $U, V \subseteq \{0, 1\}^d$, both of the same size n , and an integer $r \in \{0, 1, \dots, d\}$, are there two vectors $u \in U$ and $v \in V$ such that their inner product does not exceed r , i.e. $u \cdot v := \sum_{i=1}^d u_i \cdot v_i \leq r$?*

Lemma 7 (Abboud, Backurs, Vassilevska Williams [1]). *If Most-Orthogonal Vectors on n vectors in $\{0, 1\}^d$ can be solved in $T(n, d)$ time, then given a CNF formula on N variables and M clauses, we can compute the maximum number of satisfiable clauses (MAX-CNF-SAT), in $O(T(2^{N/2}, M) \cdot \log M)$ time.*

In our reduction we use alphabet Σ consisting of the integers from 3 to $3d + 2$. First we define two kinds of *coordinate gadgets*: $\text{CG}_1, \text{CG}_2 : \{0, 1\} \times \{1, 2, \dots, d\} \rightarrow \Sigma^*$.

$$\begin{aligned} \text{CG}_1(0, i) &= \langle 3i, 3i + 1 \rangle \\ \text{CG}_1(1, i) &= \langle 3i + 2 \rangle \\ \text{CG}_2(0, i) &= \langle 3i, 3i + 2 \rangle \\ \text{CG}_2(1, i) &= \langle 3i + 1 \rangle \end{aligned}$$

Observe that

$$\text{LCWIS}(\text{CG}_1(x, i), \text{CG}_2(y, i)) = \begin{cases} 0, & \text{if } x = 1 \text{ and } y = 1, \\ 1, & \text{otherwise.} \end{cases}$$

Now we can define two kinds of *vector gadgets*: $\text{VG}_1, \text{VG}_2 : \{0, 1\}^d \rightarrow \Sigma^*$.

$$\begin{aligned} \text{VG}_1(u_1, u_2, \dots, u_d) &= \text{CG}_1(u_1, 1) \text{CG}_1(u_2, 2) \dots \text{CG}_1(u_d, d) \\ \text{VG}_2(u_1, u_2, \dots, u_d) &= \text{CG}_2(u_1, 1) \text{CG}_2(u_2, 2) \dots \text{CG}_2(u_d, d) \end{aligned}$$

Observe that

$$\text{LCWIS}(\text{VG}_1(u), \text{VG}_2(v)) = d - (u \cdot v).$$

The next lemma helps to us combine many vector gadgets into two sequences such that computing their WLCWIS lets us find two vector gadgets with the largest LCWIS, which corresponds to the pair of most orthogonal vectors.

Lemma 8. Let $S = \{s_1, s_2, \dots, s_n\} \subseteq \Sigma^*$, $T = \{t_1, t_2, \dots, t_n\} \subseteq \Sigma^*$. Augment the alphabet Σ with four additional symbols A, B, Y, Z, such that $A < B < \sigma < Y < Z$ for all $\sigma \in \Sigma$. Denote by ℓ the maximum total weight of any sequence in $S \cup T$, and let the weights of the new symbols be $w(A) = w(Z) = \ell$ and $w(B) = w(Y) = 2\ell$. Finally, define two sequences P_1 and P_2 ,

$$P_1 = A^{2n} s_1 YB s_2 YB \dots YB s_n Z^{2n},$$

$$P_2 = (ZYBA)^n t_1 ZYBA t_2 ZYBA \dots ZYBA t_n (ZYBA)^n.$$

Then

$$\text{WLCWIS}(P_1, P_2) = \max_{1 \leq i, j \leq n} \text{WLCWIS}(s_i, t_j) + (4n - 2) \cdot \ell.$$

Proof. The proof consists of two parts, first we prove that a common weakly increasing subsequence of P_1 and P_2 with the claimed total weight exists, then we prove that there is no such subsequence with a larger total weight.

Take i and j maximizing $\text{WLCWIS}(s_i, t_j)$, denote by Q the corresponding subsequence of s_i and t_j , and consider the following sequence:

$$Q' = A^{n+j-1-(i-1)} B^{i-1} Q Y^{n-i} Z^{2n-j-(n-i)}.$$

The construction of Q' is motivated by the following intuition. We start with Q , and append at the beginning one B for each YB-fragment appearing before s_i in P_1 . There are $i - 1$ such fragments. On the other hand, the number of ZYBA-fragments appearing before t_j in P_2 equals $n + j - 1$. We devote the rightmost $i - 1$ of these fragments to find matching B symbols, and for each of the remaining fragments we take one A and append it at the beginning of Q' . The A^{2n} prefix of P_1 is long enough to match all these A symbols. In an analogous way we append Y and Z elements at the end of Q' .

Note that Q' is weakly increasing and it appears in both P_1 and P_2 as a subsequence. The total weight of Q' equals

$$\begin{aligned} \ell(n + j - 1 - (i - 1)) + 2\ell(i - 1) + \text{WLCWIS}(s_i, t_j) + 2\ell(n - i) + \ell(2n - j - (n - i)) = \\ = \ell \cdot 2n + 2\ell \cdot (n - 1) + \text{WLCWIS}(s_i, t_j) = \text{WLCWIS}(s_i, t_j) + (4n - 2) \cdot \ell. \end{aligned}$$

This finishes the first part of the proof.

Now we prove that there is no common weakly increasing subsequence of P_1 and P_2 with a larger total weight. Consider any such subsequence. Denote by c_S the number of s_i -fragments in P_1 that contribute at least one symbol to that subsequence. Analogously, denote by c_T the number of t_j -fragments in P_2 contributing at least one symbol. Finally, denote by c_A, c_B, c_Y, c_Z the number of symbols A, B, Y, Z in the subsequence, respectively. Observe that the contribution of A, B, Y, Z symbols to the total weight of the subsequence is

$$\begin{aligned} c_A \cdot w(A) + c_B \cdot w(B) + c_Y \cdot w(Y) + c_Z \cdot w(Z) = \\ = c_A \cdot \ell + c_B \cdot 2\ell + c_Y \cdot 2\ell + c_Z \cdot \ell = \\ = (c_A + c_B + c_Y + c_Z) \cdot \ell + (c_B + c_Y) \cdot \ell. \quad (9) \end{aligned}$$

Now the proof splits into two cases:

Case 1: $c_S \leq 1$ and $c_T \leq 1$. The contribution of s_i - and t_j -fragments to the total weight of the subsequence is at most $\max_{1 \leq i, j \leq n} \text{WLCWIS}(s_i, t_j)$. The remaining weight of the subsequence must come from the symbols A, B, Y, Z. Each ZYBA-fragment in P_2 can contribute at most one symbol to the, weakly increasing, subsequence. There are $3n - 1$ such fragments, therefore

$$c_A + c_B + c_Y + c_Z \leq 3n - 1.$$

Similarly, each YB-fragment in P_1 can contribute at most one symbol. There are $n - 1$ such fragments in P_2 , therefore

$$c_B + c_Y \leq n - 1.$$

Combining these two inequalities with the equation (9) we get that the contribution of A, B, Y, Z symbols to the total weight of the subsequence is at most $(3n - 1) \cdot \ell + (n - 1) \cdot \ell = (4n - 2) \cdot \ell$. This together with the contribution of the s_i - and t_j -fragments gives the required upper bound for $\text{WLCWIS}(P_1, P_2)$.

Case 2: $c_S \geq 2$ or $c_T \geq 2$. The contribution of s_i - and t_j -fragments to the total weight of the subsequence is at most $\min(c_S, c_T) \cdot \ell$. Since $\min(c_S, c_T) = c_S + c_T - \max(c_S, c_T)$ and at least one of c_S, c_T is at least 2, we can bound $\min(c_S, c_T)$ from above by $c_S + c_T - 2$. Thus the considered contribution is at most

$$((c_S - 1) + (c_T - 1)) \cdot \ell.$$

If two s_i -fragments contribute at least one symbol each, no YB-fragment between them can contribute, because the subsequence has to be weakly increasing. Therefore, the number of YB-fragments contributing is at most $n - 1 - (c_S - 1)$, thus

$$c_B + c_Y \leq n - 1 - (c_S - 1).$$

Similarly

$$c_A + c_B + c_Y + c_Z \leq 3n - 1 - (c_T - 1).$$

Using the equation (9) we get that the total weight of the subsequence is at most

$$((c_S - 1) + (c_T - 1)) \cdot \ell + (3n - 1 - (c_T - 1)) \cdot \ell + (n - 1 - (c_S - 1)) \cdot \ell,$$

which is $(4n - 2) \cdot \ell$. □

Now we are ready to show our reduction from Most-Orthogonal Vectors to LCWIS.

Lemma 10. *If LCWIS of two sequences of length n can be computed in $O(n^{2-\varepsilon})$ time, then Most-Orthogonal Vectors for n vectors in $\{0, 1\}^d$ problem can be solved in $O((dn)^{2-\varepsilon})$.*

Proof. Given two sets of vectors U and V , we use VG_1 to construct vector gadgets for vectors from set U , and VG_2 to construct vector gadgets for vectors from set V . We get two sets, each containing n vector gadgets of length at most $2d$. Now we put weight 1 for every symbol used to construct the vector gadgets, and use Lemma 8 to construct two sequences, of total weight $O(dn)$, such that their WLCWIS can be used to calculate the inner product of the pair of most orthogonal vectors. Finally, we use Lemma 5 to reduce computing the WLCWIS of these two sequences to computing LCWIS of two sequences of length $O(dn)$. □

Proof of Theorem 2. This theorem is a direct conclusion from Lemma 7 and Lemma 10. □

4 Reduction from BP-SAT to LCWIS

Abboud et al. [2] prove that if LCS for two sequences of length n can be computed in $O(n^{2-\varepsilon})$ time, then given a nondeterministic Branching Program of length T and width W on N variables it is possible to decide if there is an assignment to the variables that makes the program accept in $O(2^{(1-\varepsilon/2)N} \text{poly}(T^{\log W}))$ time. The existence of such an algorithm would have much more remarkable consequences in computational complexity than just refuting SETH. For an in-depth discussion of these consequences we refer the curious reader to [2].

The above-mentioned reduction exploits properties of LCS in three ways: when reducing LCS-Pair to LCS (Lemma 1 in [2]), when constructing the AND gadgets (Lemma 2 in [2]), and when constructing the OR gadgets (Lemma 3 in [2]).

Our techniques can be used to show a similar reduction to LCWIS. We follow the general construction of [2], and replace parts specific to LCS with LCWIS analogues. Lemma 1 in [2] is replaced with our Lemma 8. The AND gadget for LCWIS can be obtained by concatenating the corresponding sequences and using the disjoint union of their alphabets as the alphabet for concatenation. In the new alphabet all symbols from the first sequence should be made smaller than all symbols from the second sequence. To obtain the OR gadget we use the same disjoint union construction for the alphabet, and apply our Lemma 8 to resulting sequences. Finally, we follow the proof of Theorem 3 in [2] to glue these three pieces together and get the reduction from BP-SAT to LCWIS.

5 Open problems

The reductions which we use in this paper do not work with a constant size alphabet. In the case of LCS it is possible to obtain conditional quadratic time lower bounds even for binary strings [5]. It is very unlikely to obtain similar results for LCWIS given the linear time algorithm for the 3-letter alphabet [6]. However, it is an open problem to show quadratic time hardness of LCWIS for an alphabet of some larger constant size.

Even though $O(n^2)$ algorithms for LCIS and LCWIS are virtually identical, our reductions cannot be easily modified to work with LCIS. The status of LCIS is thus open. It is possible that it can be solved in strongly subquadratic time, but it is also possible that there is a reduction proving that such an algorithm is unlikely.

The existing techniques for constructing reductions appear insufficient for both of these open problems and some new ideas seem necessary.

References

- [1] ABBOUD, A., BACKURS, A., AND WILLIAMS, V. V. Tight hardness results for LCS and other sequence similarity measures. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)* (Washington, DC, USA, 2015), FOCS '15, IEEE Computer Society, pp. 59–78.
- [2] ABBOUD, A., HANSEN, T. D., WILLIAMS, V. V., AND WILLIAMS, R. Simulating branching programs with edit distance and friends: Or: a polylog shaved is a lower bound made. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing* (New York, NY, USA, 2016), STOC 2016, ACM, pp. 375–388.

- [3] ARORA, S., AND BARAK, B. *Computational Complexity: A Modern Approach*, 1st ed. Cambridge University Press, New York, NY, USA, 2009.
- [4] BACKURS, A., AND INDYK, P. Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* (New York, NY, USA, 2015), STOC '15, ACM, pp. 51–58.
- [5] BRINGMANN, K., AND KUNNEMANN, M. Quadratic conditional lower bounds for string problems and dynamic time warping. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)* (Washington, DC, USA, 2015), FOCS '15, IEEE Computer Society, pp. 79–97.
- [6] DURAJ, L. A linear algorithm for 3-letter longest common weakly increasing subsequence. *Information Processing Letters* 113, 3 (2013), 94 – 99.
- [7] KUTZ, M., BRODAL, G. S., KALIGOSI, K., AND KATRIEL, I. Faster algorithms for computing longest common increasing subsequences. *Journal of Discrete Algorithms* 9, 4 (2011), 314 – 325. Selected papers from the 17th Annual Symposium on Combinatorial Pattern Matching (CPM 2006).
- [8] MASEK, W., AND PATERSON, M. A faster algorithm computing string edit distances. *Journal of Computer and System Sciences* 20 (1980), 18–31.
- [9] SAKAI, Y. A linear space algorithm for computing a longest common increasing subsequence. *Inf. Process. Lett.* 99, 5 (Sept. 2006), 203–207.
- [10] YANG, I.-H., HUANG, C.-P., AND CHAO, K.-M. A fast algorithm for computing a longest common increasing subsequence. *Inf. Process. Lett.* 93, 5 (Mar. 2005), 249–253.